Erschienen in: Kratochvílová, Iva/Wolf, Norbert Richard (Hrsg.): Grundlagen einer sprachwissenschaftlichen Quellenkunde. - Tübingen: Narr, 2013. S. 121-134.
(Studien zur Deutschen Sprache 66)

Wilfried Schütte, IDS Mannheim

Metadaten für Gesprächsdatenbanken: ein Überblick und ihre Verwaltung in der IDS-Datenbank Gesprochenes Deutsch (DGD)

Jede Datenbank zu Text-Korpora benötigt Metadaten, ist ohne sie nur eine amorphe Sammlung sprachlicher Daten. Im Folgenden beschäftige ich mich mit Metadaten zu Gesprächsdatenbanken, also zu recherchierbaren Archiven aus Korpora natürlicher mündlicher Interaktion.[1]

Was sind Metadaten? Diese ‘Daten über Daten’ sind in erster Linie soziodemografische Daten zu Sprechereignissen und Sprechern. Im Sinne der Korpusverwaltung umfassen sie aber auch Angaben zum Speicherort und -medium von Gesprächsaufnahmen, Angaben zu technischen Aufnahmeparametern, Angaben zur Datenaufbereitung (Transkription) und Zusatzmaterialien. Das können Kommunikationsverläufe, von den Gesprächsteilnehmern benutzte Texte oder andere Medien sein. Zu vielen Feldaufnahmen existieren ethnografische Dokumente, also Memoschreiben, Feldtagebücher und andere Felddokumente wie Fotos oder Publikationen.

Wozu werden diese Metadaten benötigt? Um dem Arbeitsprozess in Korpusprojekten zu folgen: Sie dienen zunächst der Verwaltung von Dokumentationsdaten während der Korpuserstellung, sind also Werkzeuge für ein Projektmanagement und ein Steuerungsmittel für die Felderschließung. Später sind sie unabdingbar für die Recherche in Gesprächskorpora und -datenbanken: Metadaten dienen zum einen der Vorauswahl, in welchen Korpora bzw. welchen Transkripten gesucht werden soll. In einem weiteren Sinne haben sie aber auch eine ‘Filter’-Funktion: Aus der Gesamtheit der Aufnahme werden für Rechercheanfragen nur bestimmte ausgewählt. Man kann so auch virtuelle Korpora zu Recherchezwecken bestimmen – z. B. nur Aufnahmen aus Mannheim nehmen und die Verteilung sprachlicher Phänomene in diesen Aufnahmen und Transkripten mit der Verteilung in der Grundgesamtheit vergleichen. Metadaten ermöglichen so eine strukturierte Recherche, also kombinierte Suchanfragen zu Transkriptionstext und soziodemografischen Daten. Eine solche Anfrage könnte z. B. das Vorkommen des Worts *aber* im Gesprächstyp ‘Talkshow’ betreffen.[2]

---

[1] Ich danke Sylvia Dickgießer und Joachim Gasch für wichtige Hinweise hierzu.

[2] Merkel/Schmidt (2009) listen auf, welche Typen von Metadaten in online verfügbaren Korpora gesprochener Sprache erfasst werden und bei welchen dieser Angebote auch in den Metadaten recherchiert werden kann.

Es gibt eine Reihe von Metadatenschemata außerhalb des IDS:

- Dublin Core Metadata Initiative, vgl. http://dublincore.org/
- Open Language Archives (OLAC), vgl. http://www.language-archives.org/
- Text Encoding Initiative (TEI), vgl. http://www.tei-c.org/
- MPEG 7, vgl. http://www.chiariglione.org/mpeg/standards/mpeg-7/mpeg-7.htm
- ISLE Meta Data Initiative (IMDI), vgl. http://www.mpi.nl/IMDI/

In IMDI werden Metadaten für multimodale Korpora spezifiziert; diese Metadaten werden sowohl zur Katalogisierung als auch zur Information des Nutzers benutzt. Mit dem zugehörigen IMDI-BC-Browser lässt sich eine Hierarchie von Korpora und Subkorpora durchforsten. Die Basisstruktur eines IMDI-Metadaten-Satzes wird von Brugman/Broeder/Senft so beschrieben (Brugman/Broeder/Senft in URL1):

- General metadata: Name and Title of the session together with a specification where it was recorded (Location). Information on the Project and the Collector etc.
- Content part: information on what the session is about; fine-grained linguistic categorisation system for this.
- Participants part: information about the consultants whose linguistic performance is the subject of the session.
- Resources part (annotations, media files): information about the format and place of these files."

Was sind die Vor- und Nachteile dieser konkurrierenden Metadatenschemata? Ich referiere dazu eine vergleichende Bewertung (Trippel/Baumann in URL2, S. 21):

„Für die Archivierung von Ressourcen sind verschiedene Standards definiert und betrachtet worden:

- **Dublin Core:** kleinster gemeinsamer Nenner von Metadaten […], wobei der Schwerpunkt auf der Katalogisierung von Ressourcen liegt.
- **OLAC:** DC Erweiterung für mehrsprachige und vor allem auch in anderen Medien vorliegenden Ressourcen.
- **TEI:** Struktur für Metadaten für gedruckte und textuelle Medien […], wobei weder Mehrsprachigkeit noch andere Medien vorgesehen sind.

- **IMDI:** geeignetster Standard, da er die anderen Standards konzeptuell einschließt und gleichzeitig Probleme mehrsprachiger Ressourcen und verschiedener Medien berücksichtigt. Einzig die fehlenden Datenkategorien auf Annotationsebene stellen ein Problem dar, wobei aber auch in anderen Standards hierfür keine Kategorien bekannt sind."

Im Institut für Deutsche Sprache ist in den vergangenen Jahren eine neue Version der Datenbank für Gesprochenes Deutsch (DGD 2.0) entwickelt worden. Bei der Frage, welches Metadatenschema dafür verwendet werden solle, ergab sich bei der Prüfung von IMDI eine kritische Einschätzung: Zum einen enthält das IMDI-Schema, um eine Vergleichbarkeit der Daten gewährleisten und zugleich sehr heterogenen Datenbeständen gerecht werden zu können, nur eine relativ kleine Anzahl verbindlicher Informationselemente. Daneben sind in allen Abschnitten optionale 'description elements', in denen unstrukturierte Texte abgelegt werden können, sowie optionale 'keys' vorgesehen, die es verschiedenen Forschungsgruppen ermöglichen, gruppenspezifische Strukturen in das Schema zu integrieren. Dadurch wird eine große Flexibilität gewährleistet, aber auch eine Beliebigkeit gefördert, die für die Zwecke der DGD 2.0 nicht sinnvoll ist.

Zum anderen bezieht sich das Session-Konzept auf 'linguistic events' und speichert alle Daten über die sozialen Kontexte ('circumstances and conditions') dieser 'linguistic events' sowie alle Daten für die an einer 'Session' Beteiligten ('actors'), in einem Schema. Das führt zu Redundanzen in der Datenbasis, wenn mehrere 'linguistic events' in einem sozialen Kontext zu beschreiben sind und wenn einzelne Personen an mehreren dieser 'linguistic events' beteiligt waren.

Ich möchte nun darstellen, wie Metadaten in der DGD 2.0 des IDS erfasst und verwaltet werden, und stütze mich dabei auf Dickgießer/Gasch (Dickgießer/Gasch in URL 3). Die Metadatenkomponente beruht auf einem neuen Modell für die Dokumentation von Korpora der gesprochenen Sprache und umfasst 4 (XML-)Schemata. Richtlinien dabei waren:

- Unabhängigkeit von spezifischen Forschungsansätzen
- Vermittlung zwischen projektübergreifenden und projektspezifischen Anforderungen
- detaillierte Datenstruktur
- kalkulierte Redundanz
- validierbare Datenerfassung

- konsistente zentrale Datenspeicherung
- variable benutzerfreundliche Darstellung
- effektives korpusübergreifendes Retrieval
- datenschutz- und datensicherheitsgerechte Benutzerverwaltung.

Das Datenmodell für die Dokumentation von Korpora der gesprochenen Sprache sieht vier Bereiche vor, die mithilfe von (XML-)Schemata strukturiert werden: für Ereignisdaten, für ereignisübergreifende allgemeine Sprecherdaten, für Informationen über Zusatzmaterialien auf Korpusebene (z. B. Transkriptionskonventionen oder Texte, die von allen Informanten vorgelesen wurden) und für eine Korpusbeschreibung.

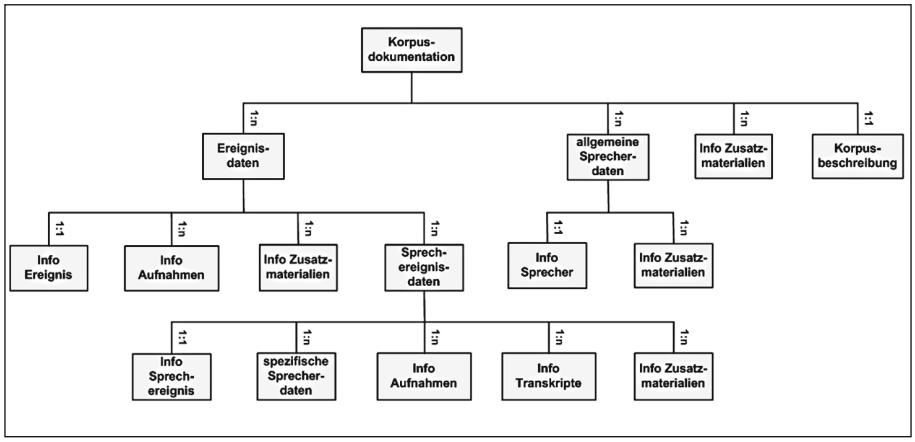


Abb. 1: IDS-Datenmodell für die Dokumentation von Korpora der gesprochenen Sprache (Dickgießer/Gasch 2011,6)

Unterschiede zwischen dem IDS-Metadatenschema für die DGD 2.0 und IMDI betreffen zum einen die Unterscheidung zwischen Ereignis und Sprechereignis. Zum anderen werden Sprecherdaten in zwei Bereichen abgelegt: ereignis- und sprechereignisspezifische Sprecherdaten im Sprechereigniskomplex des Bereichs Ereignisdaten, sprechereignis- und ereignisübergreifende Sprecherdaten im separaten Bereich für allgemeine Sprecherdaten. Die Vorteile sind ein geringes Maß an Redundanz in der Datenbasis, wenn mehrere Sprechereignisse pro Ereignis zu dokumentieren sind und Sprecher zu dokumentieren sind, die an mehreren Sprechereignissen beteiligt waren.

Bei der Metadatenverwaltung für die Korpora gesprochener Sprache im IDS wird zwischen generischen Schemata und projekt-spezifischen Subschemata unterschieden. Generische Schemata setzen Standards; sie enthalten obligato-

rische (‘mandatory’) und fakultative Komponenten, Felddefinitionen und Standardwerte; sie bilden so die Grundlage für projektspezifische Subschemata. Projekt-spezifische Subschemata übernehmen alle obligatorischen Komponenten, ergänzen sie durch eine Auswahl fakultativer Komponenten, die für das auswählende Projekt verbindlich werden. Einzelne, in den generischen Schemata vorgegebene Werte können an Projektbedürfnisse angepasst werden, indem projektspezifischer Muster spezifiziert werden, mit denen die eingegebenen Werte schon bei der Erfassung verglichen werden, und indem Feldern mit projektspezifischen Werten vorbelegt werden, u. a. in Form von Auswahllisten, wobei die Vorgaben verschiedener Projekte koordiniert werden sollten.

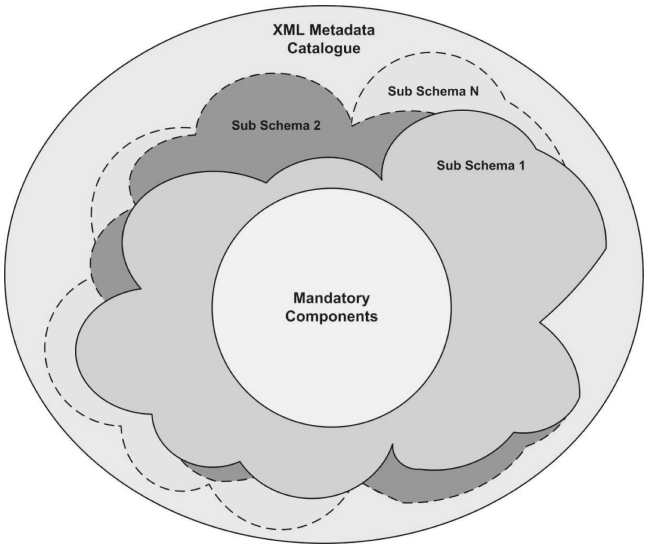


Abb. 2: Generische Schemata und projektspezifische Subschemata (Dickgießer/Gasch 2011,6)

Zur Unterscheidung zwischen ‘Ereignis’ und ‘Sprechereignis’: Als ‘Ereignis’ wird eine Phase des sozialen Geschehens verstanden, die von Beteiligten bzw. Korpusproduzenten als abgrenzbare Einheit wahrgenommen und aufgezeichnet wird. Beispielsweise gelten im Korpusprojekt ‘Deutsch heute’ (DH) mehrstündige Aufnahmesitzungen in Schulen und Volkshochschulen, die von Projektmitarbeitern geleitet wurden, als zu dokumentierende Ereignisse. Das IDS-Korpus ‘Stadtsprache Mannheim’ enthält u. a. Aufzeichnungen von Treffen sozialer Gruppen in bestimmten Stadtteilen. Jedes dieser Gruppentreffen kann als Ereignis dokumentiert werden. Ein im IDS-Korpus ‘Biographische und Reiseerzählungen’ aufgezeichnetes Ereignis wurde folgendermaßen beschrieben:

„Gemeinsames Kaffeetrinken während einer Seminarpause. Das Treffen zwischen Studentinnen und Dozentinnen wurde organisiert, um Reiseerzählungen aufzuzeichnen."

Demgegenüber ist ein 'Sprechereignis' der aufgezeichnete kommunikativer Gehalt eines Ereignisses bzw. Segmente dieses Gehalts. So gilt im Korpusprojekt 'Deutsch heute' jede Aufgabe, die im Rahmen einer mehrstündigen Aufnahmesitzung bearbeitet wurde, als Sprechereignis. Zu diesen Aufgaben gehören u. a. Bildbenennung, Verlesen einer Wortliste, Übersetzung und Interview. Im IDS-Korpus 'Stadtsprache: Mannheim' sind Aufnahmen von Gruppentreffen enthalten, bei denen z. B. Witze erzählt, Klatsch ausgetauscht und gemeinsame Unternehmungen geplant wurden. Solche kommunikativen Sequenzen können als einzelne Sprechereignisse dokumentiert werden. Das IDS-Korpus 'Elizitierte Konfliktgespräche' enthält Aufzeichnungen von Settings, in denen jeweils eine Mutter-Tochter-Dyade zwei Konfliktgespräche führte. Das Thema des ersten Gesprächs wurde von der Mutter eingebracht, das Thema des zweiten von der Tochter. Jedes dieser Gespräche gilt als ein Sprechereignis.

Im Datenmodell fungiert 'Ereignis' als Startknoten eines generischen (XML-) Schemas für unterschiedliche Angaben – über Aufzeichnungsobjekte (Ereignis, Sprechereignis, Sprecher), über Korpusbestandteile (Audioaufnahmen, Videoaufnahmen, Transkripte, Zusatzmaterialien auf Ereignis- und Sprechereignisebene) und für die Dokumentationsgeschichte.

Obligatorische Komponenten sind in allen projektspezifischen Subschemata zu berücksichtigen. Fakultative Komponenten: stehen zur Wahl – wenn sie verwendet werden, müssen alle Kennungsfelder und alle Felder die ein Fragezeichen-Symbol enthalten bearbeitet werden. Eingaben für fehlende Daten in Feldern mit Fragezeichen sind standardisiert. 'Nicht dokumentiert' heißt: Es kann ein Datum geben, das bei der Datenerfassung jedoch nicht bekannt ist. 'Nicht vorhanden' heißt: Es gibt kein Datum. Einige Komponenten des Schemas sind iterativ, sie können bei der Datenerfassung vervielfältigt werden.

Wichtig ist auch eine Dokumentation der Quellaufnahmen. Darunter werden Rohdaten verstanden, als Originalaufnahmen von Ereignissen oder Aufnahmen, die für die dokumentierende Stelle Originalcharakter haben. Dazu gibt es unterschiedliche Typen: Audioaufnahme, Videoaufnahme oder Tonkopie von Videoaufnahmen. Das Feld 'Schutzbedürftige_Daten' wird belegt, wenn die Quellaufnahme Daten enthält, die nach dem Willen der Urheber und aus datenschutzrechtlichen Gründen Außenstehenden nicht kenntlich werden dürfen, wie z. B. persönliche Sprecherdaten. Hier wird auch notiert, ob eine vollständige oder eine unvollständige Aufnahme eines Ereignisses dokumentiert wird.

Als ‘Zusatzmaterial’ verzeichnet werden Dokumente, die zusätzlich zu Quellaufnahmen vorhanden sein können, z. B. Reiseberichte der Aufnahmeleiter, Protokolle von Aufnahmesitzungen, Fotos von Aufnahmeorten oder Notizen zu einer Sitzordnung. Dieser Komplex ist fakultativ und iterativ (für den Fall, dass mehrere Dokumente zu einem Ereignis zu beschreiben sind).

Die Basisdaten eines Sprechereignisses umfassen einen linguistisch unspezifischen Feldnamen ‘Art’, der hier verwendet wird anstelle von Kategorien wie ‘Textsorte’, ‘Texttyp’, ‘Interaktionstyp’, ‘Gesprächstyp’, ‘Diskurstyp’, ‘Genre’, ‘Gattung’, die aus verschiedenen Forschungsansätzen stammen, um Daten aus allen Bereichen aufnehmen zu können. Angaben können hier ‘Erzählung’, ‘Rede’, ‘Anleitung’, ‘Beschreibung’, ‘Benennung’, ‘Übersetzung’, ‘Interview’, ‘Beratung’, ‘Diskussion’, ‘Begrüßung’ etc. sein, evtl. mehrfach. Im nächsten Schritt wird die Anzahl der Sprecher notiert, wobei verbal beteiligte Forscher/Aufnahmeleiter mitgezählt werden sollten, was man dann im Feld ‘Forscherbeteiligung’ verdeutlichen kann. Dort sind die Werte ‘Verbal beteiligt’, ‘Nicht verbal beteiligt’ und ‘Nicht vorhanden’ (für ‘Forscher nicht anwesend’) vorgesehen. ‘Elizitierung’ ist eine Technik zur Erhebung sprachlicher Daten, bei der die Informanten systematisch zu Äußerungen veranlasst werden. Vorgesehen sind hier die Werte ‘Elizitiert’ und ‘Nicht elizitiert’. ‘Mediale_Realisierung’ steht für den jeweiligen Kommunikationskanal (wie z. B. ‘Face to Face’, ‘Telefon’, ‘Hörfunk’). Für das Feld ‘Öffentlichkeitsgrad’ werden die Werte ‘Öffentlich’ und ‘Nicht öffentlich’ bereitgestellt. Über Instruktionen und ggf. auch über Materialien, die den Sprechern zur Lösung bestimmter Aufgaben vorgelegt wurden, kann man im Feld ‘Vorgaben’ informieren. Die Position eines Sprechereignisses im Ereignis kann relevant sein, wenn Segmente des aufgezeichneten kommunikativen Gehalts eines Ereignisses betrachtet werden. In solchen Fällen kann man hier die Zusammenhänge beschreiben. Eine mögliche Positionsbeschreibung wäre: ‘Beginnt unmittelbar nach der Begrüßung der Beteiligten und endet vor der ersten längeren Pause’. Im Feld ‘Sprachen’ sind die im Sprechereignis verwendeten Sprachen zu verzeichnen.

Die Basisdaten zur Transkription und Annotation betreffen u. a.die ‘Spezifikation’ als Charakterisierung der Annotation (Gegenstand, Umschrift, Reichweite), die ‘Konventionen’ (z. B. ‘projektspezifisch’, ‘DIDA, Version vom Januar 2001’, ‘GAT-2’) und das ‘Zeicheninventar’ (z. B. IPA, spezifische Alphabete). Das Alignment von Transkripten betrifft die ‘Text-Ton-Synchronisation’, also die Koppelung von Aufnahmen und Transkripten auf Phon-, Phonem-, Wort- oder Phrasenbasis. Dabei werden Zeitmarken Transkriptsegmenten zugeordnet. Das kann manuell oder nach einem automatischen Alignment-Ver-

fahren geschehen und wird entsprechend im Feld ‘Verfahren_Instrumente’ notiert, gegebenenfalls unter Angabe der verwendeten Software.

Sprecherdaten werden u. a. ereignis- und sprechereignisübergreifend dokumentiert. Dazu dient im Datenmodell (s. Abb. 1) die Kategorie ‘Sprecher’ als Startknoten eines (XML-)Schemas mit Informationen zu Angaben über den jeweiligen Sprecher (Basisdaten, Ortsdaten, Sprachdaten), über Beziehungen dieses Sprechers zu anderen Sprechern und sonstige Bezugspersonen des Sprechers, über Vereinbarungen zu Datenschutz und Nutzungsrechten und über Zusatzmaterial auf Sprecherebene sowie eine Dokumentationsgeschichte. Wichtig ist dabei die Rechteverwaltung: Welche rechtlichen Aspekte der Datenerhebung sowie rechtsrelevante Vereinbarungen mit Sprechern und ggf. auch Bezugspersonen über Schutz und Verwendung ihrer Daten sind für die Speicherung und Weiterverarbeitung der Daten zu beachten? Aus welchen Quellen stammen die erhobenen personenbezogenen Daten? Stammen sie von den Sprechern, aus einer Befragung von Bezugspersonen oder aus einer Auswertung schriftlicher Quellen? Sind nur besondere Arten oder alle personenbezogenen Daten zu schützen? Unter ‘Datenschutzvereinbarungen’ wird notiert, welche Vereinbarungen mit Sprechern und Bezugspersonen über den Schutz der Daten getroffen wurden. Solche Vereinbarungen können vorsehen, dass diese Daten nur im Rahmen des erhebenden Projekts verwendet und danach gelöscht werden oder auf bestimmten Wegen für bestimmte Zwecke an Dritte weitergegeben werden dürfen. Die Zustimmung der Sprecher zu den Aufnahmen ist wesentliche Voraussetzung für die Verwendung von Aufnahmen und Transkripten; daher muss festgehalten werden, ob die Sprecher über den Zweck der Aufnahmen informiert wurden, wenn ja wann. In welcher Form – schriftlich oder (aus welchen Gründen?) mündlich – haben sie den Aufnahmen zugestimmt? Nutzungsrechte an Korpusbestandteilen betreffen u. a. die wissenschaftliche Auswertung im Daten erhebenden Projekt oder eine Veröffentlichung im Internet.

Das generische Schema für die Dokumentation von Zusatzmaterial auf der Korpusebene bezieht sich auf Dokumente, die zusätzlich zu Aufnahmen, Transkripten und Zusatzmaterialien auf Ereignis-, Sprechereignis- und Sprecherebene vorhanden sein können, z. B. Transkriptionskonventionen, Interviewleitfaden, Wortlisten, verschiedene Varianten der Wenkersätze, ggf. auch Spezifikationen für die Validierung von Korpusdaten und Dokumente, die die Ergebnisse solcher Qualitätsprüfungen enthalten.

Das generische Schema für die Korpusbeschreibung soll einen systematischen Überblick über die Erstellung, die Zusammensetzung ('Aufzeichnungsobjekte'), den Bearbeitungsstand und die Verwaltung eines Korpus ermöglichen. Das 'Erstellungsprojekt' ist das Projekt, das ein Korpus aufgebaut hat. 'Korpusbestandteile' sind Quellaufnahmen von Ereignissen, sprechereignisspezifische Aufnahmen, Transkripte und Zusatzmaterial auf Ereignis-, Sprechereignis-, Sprecher- und Korpusebene.

Für das IDS-Projekt 'Forschungs- und Lehrkorpus gesprochenes Deutsch' (FOLK, agd.ids-mannheim.de/html/folk.shtml) wurden die Vorgaben des generischen Dokumentationsschemas nach spezifischen Projektbedürfnissen kondensiert und ausgewählt. Bei dieser Dokumentation musste zwischen forschungsethischen Aspekten, insbesondere dem Informantenschutz (Zusicherung von Anonymität und Nicht-Identifizierbarkeit als Einzelperson) und Forschungsinteressen, insbesondere dem dialektologischen Interesse an einer präzisen Dokumentation der Sprachbiografie, abgewogen werden. Der Datenschutz erfordert, dass Metadaten für die gesprächsanalytische Nutzung stärker gefiltert bzw. ausgedünnt werden als etwa für eine dialektologische Nutzung, die präzise Angaben zur Sprachbiografie von Sprechern verlangt. Die Daten in FOLK werden generell so maskiert, dass eine Ermittlung personenbezogener Daten weitestgehend unmöglich wird. Das heißt, Personennamen und Ortsnamen werden in den Audiodaten verrauscht und in den Transkripten durch Pseudonyme ersetzt. Personenbezogene Angaben in den Metadaten für Sprecher und Kommunikationen werden nur so präzise festgehalten, wie es für eine Verwendung der Daten in einem gesprächsanalytischen Kontext notwendig ist. D. h. insbesondere, dass Geburtsdaten nur auf das Jahr genau, Ortsdaten nur (sprach)regionengenau (z. B. 'obersächsischer Sprachraum') festgehalten werden. Zudem ist die FOLK-Dokumentation für die projektinterne Nutzung und für die externe Datenbankrecherche unterschiedlich umfangreich; in der externen Ansicht der Datenbank werden nur Ausschnitte aus den korpusspezifischen Dokumenten angezeigt.

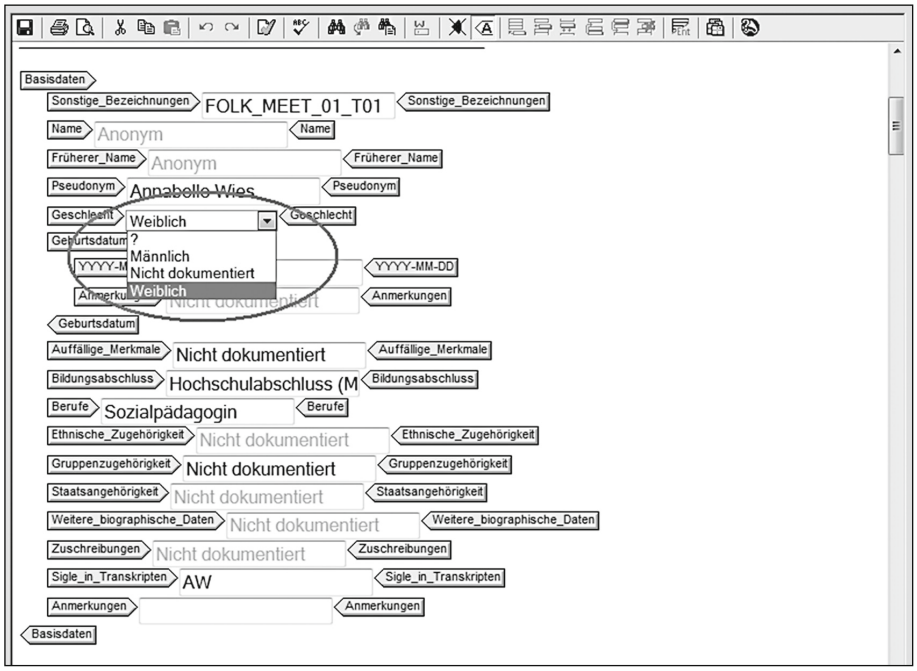


Abb. 3: Metadateneingabe im FOLK-Projekt

Das Bildschirmfoto zeigt einen Teil der Erfassung der Sprecherdokumentation. Eingetragen werden u. a. das Pseudonym der Sprecherin, das auch in den Transkripten verwendet wird, und ihr Geschlecht mit einer Ausklappliste.

Die Metadaten zu den IDS-Korpora werden u. a. in der Datenbank für Gesprochenes Deutsch (DGD), Version 2.0, publiziert. Für die DGD ist seit Februar 2012 eine Testversion online unter http://dgd.ids-mannheim.de verfügbar. Sie enthält detaillierte Beschreibungen der Korpora (Korpusmetadaten) unter ‘Korpora > Korpusbeschreibungen’. Man kann in den Ereignis- oder Sprecherdokumentationen, Transkripten und Zusatzmaterialien browsen und Ereignis- oder Sprecherdokumentationen sowie Transkripte über die entsprechenden Unterpunkte im Menü ‘Recherche > Volltext’ durchsuchen.

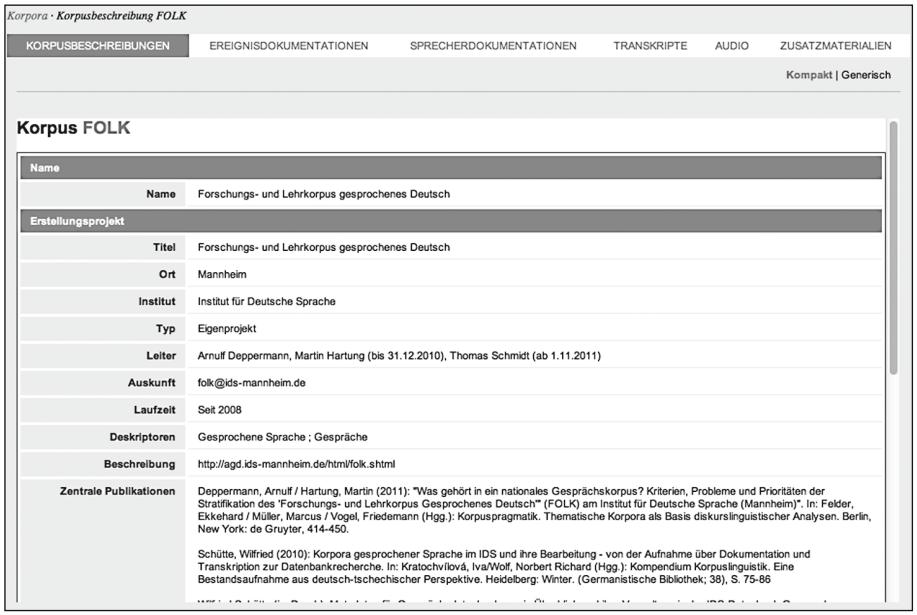


Abb. 4: DGD-Korpusbeschreibung

Als Beispiel für eine DGD-Korpusbeschreibung zeigt dieses Bildschirmfoto den Anfang der Beschreibung zum Korpus FOLK in der Kompakt-Ansicht.

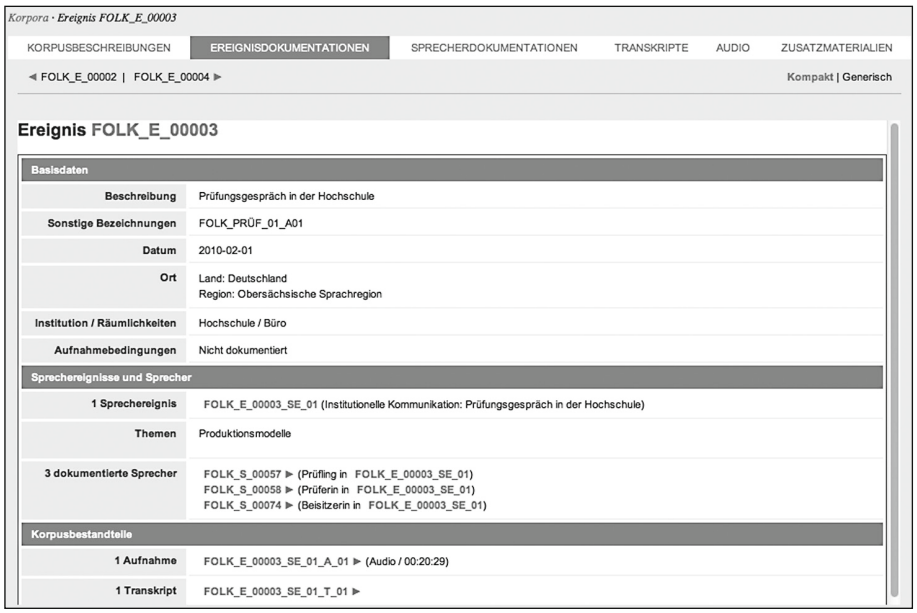
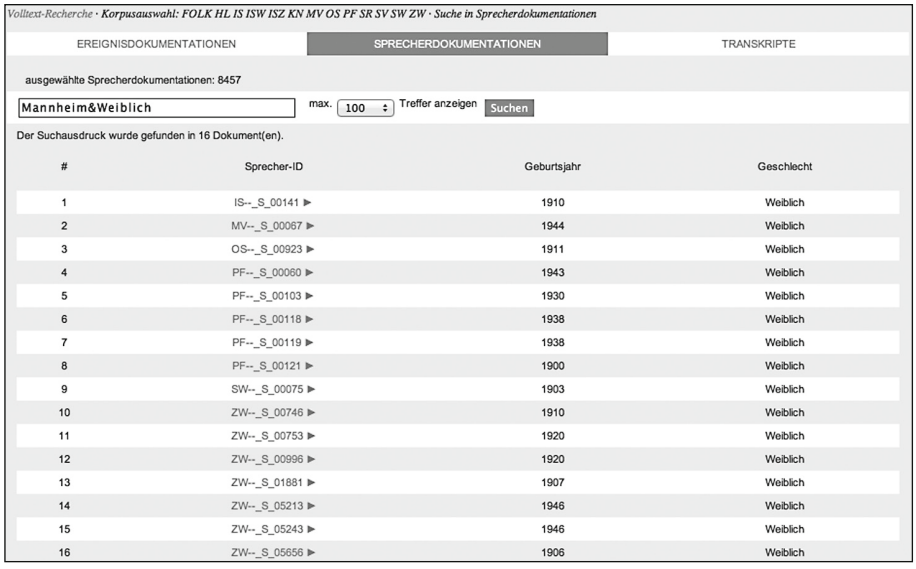


Abb. 5: DGD-Ereignisdokumentation

Als Beispiel für eine DGD-Ereignisdokumentation zeigt dieses Bildschirmfoto die Metadaten zu einem Prüfungsgespräch in der Hochschule (Kennung ‘FOLK_E_00003’), wiederum in der Kompakt-Ansicht.

Für die Volltext-Suche in Metadaten und Transkripten stehen einige Suchoptionen und -Operatoren zur Verfügung. Einfache Suchoperatoren sind ‘Wildcards’, ‘AND’, ‘OR’ und ‘NOT’. Komplexe Suchoperatoren sind:

- ‘FUZZY’ für eine Erweiterung der Suche um ähnlich geschriebene Wörter (womit auch Belege mit Tippfehlern erfasst können)
- ‘NEAR’ für eine Wortabstandssuche
- ‘SOUNDEX’ (experimentell) nach ähnlich ausgesprochenen Wörtern
- ‘STEM’, der Abfrageergebnisse mit den gleichen linguistischen Wurzeln wie der Suchausdruck liefert (Lemma-Suche), und
- ‘THRESHOLD’ für eine Suche nach Dokumenten, in denen die Gesamtzahl der Vorkommen des Suchausdrucks einen angegebenen Schwellenwert überschreitet.


Volltext-Recherche · Korpusauswahl: FOLK HL IS ISW ISZ KN MV OS PF SR SV SW ZW · Suche in Sprecherdokumentationen

EREIGNISDOKUMENTATIONEN | SPRECHERDOKUMENTATIONEN | TRANSKRIPTE

ausgewählte Sprecherdokumentationen: 8457

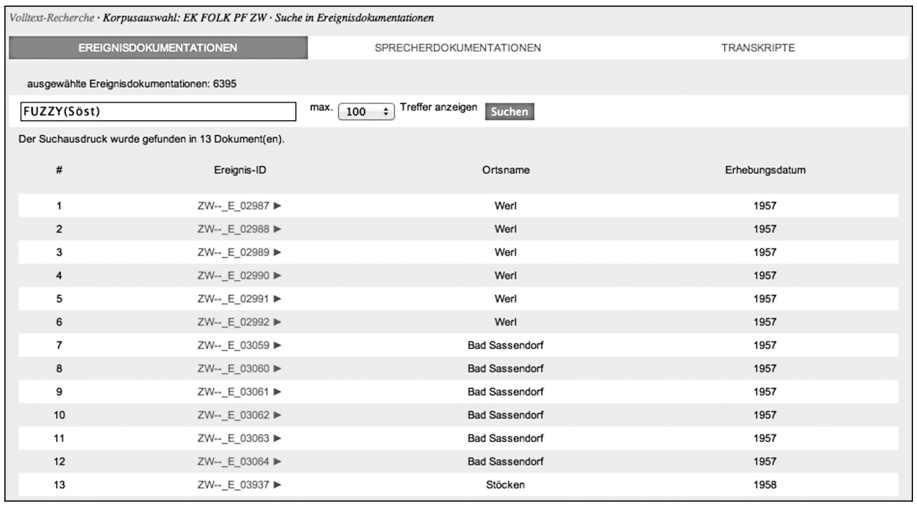
Mannheim&Weiblich max. 100 Treffer anzeigen Suchen

Der Suchausdruck wurde gefunden in 16 Dokument(en).

#	Sprecher-ID	Geburtsjahr	Geschlecht
1	IS--_S_00141 ▶	1910	Weiblich
2	MV--_S_00067 ▶	1944	Weiblich
3	OS--_S_00923 ▶	1911	Weiblich
4	PF--_S_00060 ▶	1943	Weiblich
5	PF--_S_00103 ▶	1930	Weiblich
6	PF--_S_00118 ▶	1938	Weiblich
7	PF--_S_00119 ▶	1938	Weiblich
8	PF--_S_00121 ▶	1900	Weiblich
9	SW--_S_00075 ▶	1903	Weiblich
10	ZW--_S_00746 ▶	1910	Weiblich
11	ZW--_S_00753 ▶	1920	Weiblich
12	ZW--_S_00996 ▶	1920	Weiblich
13	ZW--_S_01881 ▶	1907	Weiblich
14	ZW--_S_05213 ▶	1946	Weiblich
15	ZW--_S_05243 ▶	1946	Weiblich
16	ZW--_S_05656 ▶	1906	Weiblich

Abb. 6: Suche in Sprecherdokumentationen

Dieses Bildschirmfoto zeigt beispielhaft für die Verwendung des AND-Operators das Ergebnis einer Suche nach Sprecherinnen in Mannheim (‘Mannheim&Weiblich’) in den 8457 Sprecherdokumentationen aller derzeit verfügbaren Korpora. Der Suchausdruck wurde in 16 Dokumenten gefunden.



Volltext-Recherche · Korpusauswahl: EK FOLK PF ZW · Suche in Ereignisdokumentationen

EREIGNISDOKUMENTATIONEN | SPRECHERDOKUMENTATIONEN | TRANSKRIPTE

ausgewählte Ereignisdokumentationen: 6395

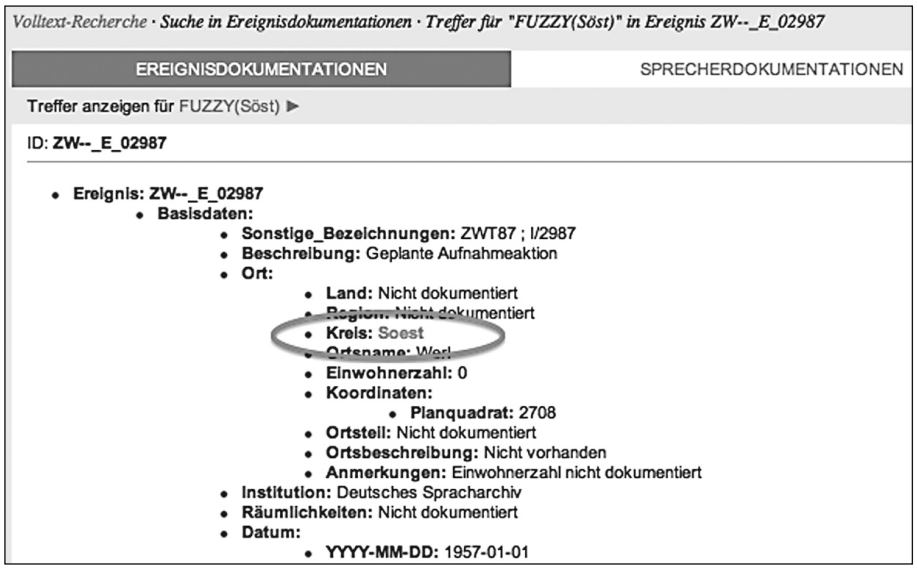
FUZZY(Söst) max. 100 Treffer anzeigen Suchen

Der Suchausdruck wurde gefunden in 13 Dokument(en).

#	Ereignis-ID	Ortsname	Erhebungsdatum
1	ZW--_E_02987 ►	Werl	1957
2	ZW--_E_02988 ►	Werl	1957
3	ZW--_E_02989 ►	Werl	1957
4	ZW--_E_02990 ►	Werl	1957
5	ZW--_E_02991 ►	Werl	1957
6	ZW--_E_02992 ►	Werl	1957
7	ZW--_E_03059 ►	Bad Sassendorf	1957
8	ZW--_E_03060 ►	Bad Sassendorf	1957
9	ZW--_E_03061 ►	Bad Sassendorf	1957
10	ZW--_E_03062 ►	Bad Sassendorf	1957
11	ZW--_E_03063 ►	Bad Sassendorf	1957
12	ZW--_E_03064 ►	Bad Sassendorf	1957
13	ZW--_E_03937 ►	Stöcken	1958

Abb. 7: FUZZY-Suche in Ereignisdokumentationen

Abschließend zeigt dieses Bildschirmfoto eine Recherche mit dem komplexen FUZZY-Operator nach ähnlich bzw. abweichend geschriebenen Wörtern mit dem hier bewusst vorgenommenen Rechtschreibfehler *Söst* (FUZZY(Söst)) statt der korrekten Schreibweise *Soest* für die westfälische Stadt. Dieses Mal wurde nur in ausgewählten Korpora gesucht, der Suchausdruck wurde in 13 Dokumenten gefunden, beispielsweise im Ereignis ZW--_E_02987 aus dem Zwirner-Korpus, dort in der Angabe des Kreises orthografisch richtig notiert:



Volltext-Recherche · Suche in Ereignisdokumentationen · Treffer für "FUZZY(Söst)" in Ereignis ZW--_E_02987

EREIGNISDOKUMENTATIONEN | SPRECHERDOKUMENTATIONEN

Treffer anzeigen für FUZZY(Söst) ►

ID: **ZW--_E_02987**

- **Ereignis: ZW--_E_02987**
  - **Basisdaten:**
    - **Sonstige_Bezeichnungen:** ZWT87 ; I/2987
    - **Beschreibung:** Geplante Aufnahmeaktion
    - **Ort:**
      - **Land:** Nicht dokumentiert
      - **Region:** Nicht dokumentiert
      - **Kreis:** Soest
      - **Ortsname:** Werl
      - **Einwohnerzahl:** 0
      - **Koordinaten:**
        - **Planquadrat:** 2708
      - **Ortsteil:** Nicht dokumentiert
      - **Ortsbeschreibung:** Nicht vorhanden
      - **Anmerkungen:** Einwohnerzahl nicht dokumentiert
    - **Institution:** Deutsches Spracharchiv
    - **Räumlichkeiten:** Nicht dokumentiert
    - **Datum:**
      - **YYYY-MM-DD:** 1957-01-01

Abb. 8: Treffer für die Suche in Ereignisdokumentationen

Die DGD 2.0 soll in den nächsten Jahren ausgebaut werden, so dass auch eine strukturierte Recherche in den Transkripten in Verbindung mit einer Vorauswahl anhand der Ereignis- und Sprecherdokumentationen möglich sein wird.

Literatur

Merkel, Silke/Schmidt, Thomas (2009): Korpora gesprochener Sprache im Netz – eine Umschau. In: Gesprächsforschung – Online-Zeitschrift zur verbalen Interaktion 10, S. 70–93. (www.gespraechsforschung-ozs.de/heft2009/px-merkel.pdf).

URL 1: Brugman, Hennie/Daan Broeder/Gunter Senft (2003): Documentation of Languages and Archiving of Language Data at the Max Planck Institute for Psycholinguistics in Nijmegen. Paper presented at the "Ringvorlesung Bedrohte Sprachen". Sprachenwert – Dokumentation – Revitalisierung. Fakultät für Linguistik und Literaturwissenschaft Universität Bielefeld (05.02.2003), www.mpi.nl/IMDI/documents/articles/BI-EL-PaperA2.pdf [29.03.2012].

URL 2: Trippel, Thorsten/Tanja Baumann (2003): Metadaten für Multimodale Korpora: Verwendung im Modelex-Projekt. Technisches Dokument 4, Universität Bielefeld (November 2003), http://www.spectrum.unibielefeld.de/modelex/publication/techdoc/modelex_techrep4/metadata_techdoc_rev2.0.pdf [29.03.2012].

URL 3: Dickgießer, Sylvia unter Mitarbeit von Joachim Gasch (2011): Metadatenschema in der Datenbank für Gesprochenes Deutsch (DGD 2.0) (01.07.2011), agd. ids mannheim.de/pdf/metadatenschemata_DGD_2.0_2011-07-01.pdf [29.03.2012].